

Development of Krylov and AMG linear solvers for large-scale sparse matrices on GPUs

Bo Yang, Hui Liu, Zhangxin Chen

Department of Chemical and Petroleum Engineering
University of Calgary
Calgary, Canada
{yang6, hui.j.liu, zhachen}@ucalgary.ca

Abstract. This research introduce our work on developing Krylov subspace and AMG solvers on NVIDIA GPUs. As SpMV is a crucial part for these iterative methods, SpMV algorithms for single GPU and multiple GPUs are implemented. A HEC matrix format and a communication mechanism are established. And also, a set of specific algorithms for solving preconditioned systems in parallel environments are designed, including ILU(k), RAS and parallel triangular solvers. Based on these work, several Krylov solvers and AMG solvers are developed. According to numerical experiments, favorable acceleration performance is acquired from our Krylov solver and AMG solver under various parameter conditions.

Keywords: Krylov subspace, GPU, SpMV, ILU, RAS, GMRES, AMG

1 Introduction

Iterative algorithms have widely applications in kinds of scientific computing fields, such as the reservoir simulation [1]. For large-scale sparse linear systems, Krylov subspace and AMG algorithms are commonly used. Krylov subspace algorithms include the GMRES (Generalized Minimal Residual), CG(Conjugate Gradient) and BiCGSTAB (Biconjugate Gradient Stabilized), etc. These algorithms are available to general matrices [2,3]. Preconditioners are always employed to optimize the performance of an iterative algorithm and many efficient preconditioners have been developed [4,5,6,20]. We have developed the Krylov subspace algorithms with ILU preconditioners. Many researchers have devoted their efforts into designing AMG solvers which is specific for symmetric positive definite matrices. Ruge and Stüben designed the RS (Ruge-Stüben) coarsening strategy and developed a classical AMG solver which is the foundation of developing other AMG solvers [7,8,9,10,11]. The parallel coarsening strategy CLJP was proposed by Luby, Jones and Plassmann [12,13]. We have also developed the AMG algorithm with a series of smoothers, coarsening operators and prolongation operators.

GPU(Graphics Processing Unit) computing emerges as an acceleration technique for image displaying. However, it has more and more utility in other sci-

entific computing disciplines. Zhang et al. completed some professional performance analysis about GPUs [14]. A NVIDIA Tesla K40 which has 2880 CUDA cores and a peak performance of 1.43 TFlops (Base Clocks) in double precision has greater performance than an Intel Core i7-5960X with 8 cores and 16 threads which has a typical peak performance of 385 GFlops [15,16]. A NVIDIA Tesla K40 also has 288 G/sec memory speed which is much faster than the speed 68 GB/s of an Intel Core i7-5960X [15,16]. As GPU has great priority in parallel computing, we have designed and developed our iterative algorithms on GPUs.

SpMV (sparse matrix-vector multiplication) is a core part for iterative algorithms. For a large and sparse matrix, it is necessary to partition it into sub matrices for GPU computation. The METIS partition method is adopted in our algorithms [17]. Because data communication is unavoidable for SpMV implementation on multiple GPUs, we have designed a specific communication mechanism for partition matrices to share vector data among different GPUs. In order to make full use of the characteristic of GPU memory access, we adopted a HEC matrix format which is more friendly to the SpMV algorithm. A NVIDIA GPU platform provides high parallel capability depending on its hundreds of fine CUDA cores. An algorithm must be designed as a parallel algorithm to run on the CUDA cores. RAS (Restricted Additive Schwarz) proposed by Cai et al. is adopted in our algorithms to improve the parallel structure of a preconditioner matrix [18]. Because the ILU preconditioners and AMG smoothers all need to solve triangular systems, we implemented a parallel triangular solver on GPUs [19]. It is based on the level schedule method [2,21]. In this research, we designed a set of numerical experiments to test our algorithms from different aspects. The experiment results and analysis are given in the experiment section.

The layout of this paper is presented as follows: In §2, the matrix format, SpMV, vector operations, ILU (k), RAS, parallel triangular solver, Krylov subspace algorithms and AMG algorithms are introduced. In §3, the numerical experiments are presented and analyzed. In §4, conclusions are given.

2 GPU Computation

2.1 Matrix Format

Several matrix formats are presented in this section. They are ELL, HYB and HEC. The ELL format is provided in ELLPACK [22]. Figure 1 shows the ELL's structure consisting of two parts. We can see the two parts are both regular and have the same dimensions. Regular storage has a high speed for data access. However, it is not wise to store a large-scale sparse matrix in such a format as lots of storage spaces are always wasted. For instance, if there are a large number of nonzero entries in one row, the other rows must maintain the same size of entries most of which are zero. In order to make the limited memory space be used efficiently, N. Bell and M. Garland suggested a hybrid matrix format named HYB (Hybrid of ELL and COO). An original matrix is split into two parts. One part is regular and the remain part is irregular. The COO format is used to store the irregular part. It has three one-dimensional arrays illustrated

in Figure 2. The HYB format has good average performance. In our research, we adopt another hybrid format called HEC which saved the irregular part in a CSR format shown in Figure 3. It also contains three one-dimensional arrays. Ap is used for storing the start position of each row. Ax and Aj have the same length and used for storing the entry data and column indices, respectively.

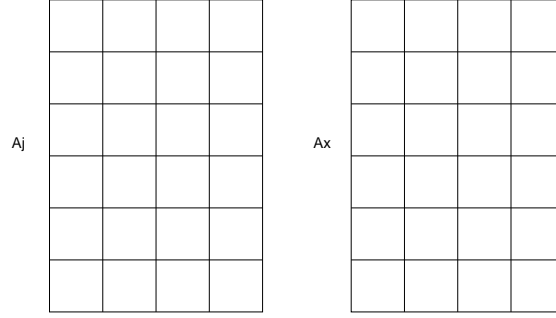


Fig. 1. ELL matrix format

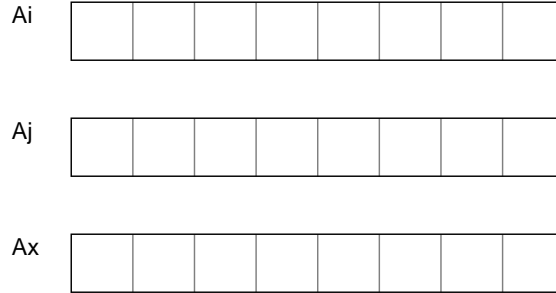
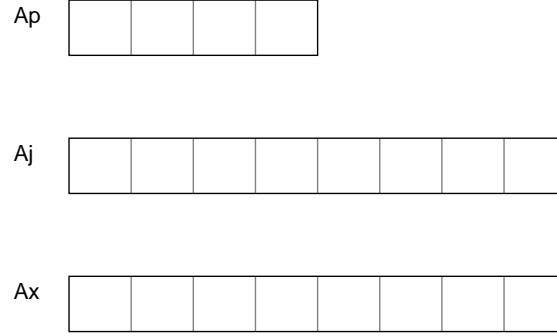


Fig. 2. COO matrix format

According to the mathematic method of SpMV, it is always calculated based on the column vectors. This can be explained by equation (1). Thereby, it's better for us to store the entries in the computer column by column. The GPU architecture provides a wrap concept to execute CUDA cores. That means 32 threads are bounded to be executed together. So the stride of the ELL part should be a multiple of 32 to acquire enhanced parallel performance. In our algorithms, we set it as 256 or other multiples. Another problem is how to decide the boundary between the ELL and CSR. We use a recommended value 20 whose theoretical explanation are introduced in [23].

**Fig. 3.** CSR matrix format

$$\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = x_1 \begin{pmatrix} A_{11} \\ A_{21} \\ \vdots \\ A_{n1} \end{pmatrix} + x_2 \begin{pmatrix} A_{12} \\ A_{22} \\ \vdots \\ A_{n2} \end{pmatrix} + \cdots + x_n \begin{pmatrix} A_{1n} \\ A_{2n} \\ \vdots \\ A_{nn} \end{pmatrix} \quad (1)$$

2.2 SpMV Algorithm

Based on the HEC matrix format, the SpMV algorithm is designed as two parts apparently. As a GPU executes hundreds of CUDA cores simultaneously, a parallel algorithm can be implemented with each CUDA core computing a row. The ELL part has high efficient and is performed firstly. Algorithm 1 gives the SpMV algorithm. This algorithm runs well on a single GPU. However, it is not suitable for multiple GPUs. Multiple GPUs bring stronger parallel computing capability but import extra data communication. We need to partition the original matrix into partition matrices first.

Algorithm 1 Sparse matrix-vector multiplication

```

1: for i = 1: n do                                     ▷ ELL
2:   Calculate the i-th row of ELL matrix;                ▷ one CUDA core
3: end for
4:
5: for i = 1: n do                                     ▷ CSR
6:   Calculate the i-th row of CSR matrix;                ▷ one CUDA core
7: end for

```

If a matrix has a regular structure. For instance, it is derived from the FDM (Finite Difference Method) or FVM (Finite Volume Method). A sequence partition method can be used. But if it is irregular structure which is often derived

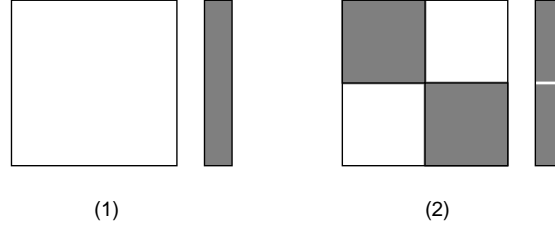


Fig. 4. Matrix and vector partition

from the FEM (Finite Element Method) or FVM. A specific partition method should be used. We select a quasi-optimal partition method METIS to complete the matrix partition. During the partition process, the rows of the matrix are switched first and all the nonzero entries are put along the diagonal as close as possible. Then the pivot blocks have most of the nonzero entries and the communication cost between any two partition matrices is reduced; see Figure 4.

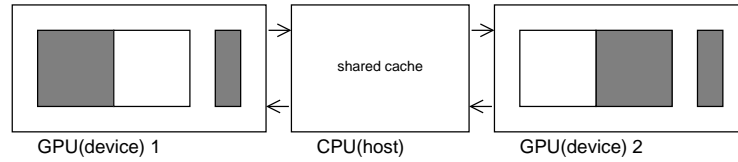


Fig. 5. Vector communication

The vector is also partitioned into segments. Each pair of a partition matrix and a segment vector is distributed onto a GPU. Although most of the nonzero entries are concentrated at the pivot block, there are still some sparse nonzero entries outside it, for which a segment vector can not provide a corresponding element to complete multiplication. Thus, the necessary communication is unavoidable. We establish a shared cache for communication. The cache is located on the CPU (host). It receives all the communication data from each GPU (device) and then sends the data to needed GPU. As we have used partition method to reduce the communication load, this mechanism is reasonable.

2.3 Vector Operations

Vector operations are necessary for developing iterative algorithms. They can be categorized into some categories by equation (2) to equation (6). Some of them are linear combinations of vectors. Some of them are about dot products. As two vectors are operated by one-to-one correspondence of elements, it is easy to design parallel algorithms for them. First, vectors are divided into segments. Then each pair of segments are distributed onto a GPU. All the sub results are

sent back to CPU after tasks are finished on GPUs. No communication cost is needed during the vector operations. A schematic is shown by Figure 6.

$$\mathbf{y} = \alpha A\mathbf{x} + \beta \mathbf{y} \quad (2)$$

$$\mathbf{y} = \alpha \mathbf{x} + \beta \mathbf{y} \quad (3)$$

$$\mathbf{z} = \alpha \mathbf{x} + \beta \mathbf{y} \quad (4)$$

$$a = \langle \mathbf{x}, \mathbf{y} \rangle \quad (5)$$

$$r = \|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} \quad (6)$$

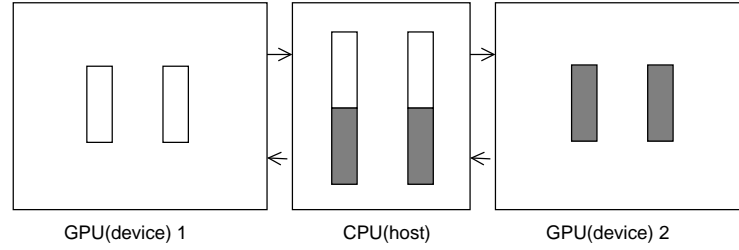


Fig. 6. Vector operations

2.4 ILU(k)

A preconditioner system is expressed as equation (7). M is the preconditioner matrix which is factorized from the original matrix A . The ILU is a commonly used preconditioner. It means M can be factorized into one lower triangular matrix L and an upper triangular matrix U , as shown by equation (8). The matrix A and LU are stored in the same memory space in the program implementation. In other words, L is stored in the low triangular part and U is stored in the upper triangular part. A level k can be used to control the factorization process. Only the entry positions meeting the requirement are allowed to have nonzero entries in the result pattern. The requirement condition is described by equation (9) and equation (10) [2].

$$M\mathbf{x} = \mathbf{y} \quad (7)$$

where

- M : the preconditioner matrix

- \mathbf{x} : the unknown vector
- \mathbf{y} : the right hand side vector

$$M = LU \quad (8)$$

$$L_{ij} = \begin{cases} 0, & (i, j) \in P \\ \infty, & (i, j) \notin P. \end{cases} \quad (9)$$

$$L_{ij} = \min\{L_{ij}, L_{ip} + L_{pj} + 1\}. \quad (10)$$

Equation 9 gives an initial level for each entry A_{ij} . P is the nonzero pattern of A . So if A_{ij} is zero, its level L_{ij} is infinite; otherwise, L_{ij} is zero. Equation (9) provides an updated algorithm for levels. This update process are executed at each loop of ILU(k) algorithm and only the satisfactory entry positions have nonzero values in the final factorization pattern. The Algorithm 2 details a complete ILU(k) procedure.

Algorithm 2 ILU(k) factorization

```

1: For all nonzero entries in nonzero pattern  $P$ , define  $L_{ij} = 0$ 
2: for  $i = 2 : n$  do
3:   for  $p = 1 : i - 1$  &  $L_{ip} \leq k$  do
4:      $A_{ip} = A_{ip}/A_{pp}$ 
5:     for  $j = p + 1 : n$  do
6:        $A_{ij} = A_{ij} - A_{ip}A_{pj}$ 
7:        $L_{ij} = \min\{L_{ij}, L_{ip} + L_{pj} + 1\}$ 
8:     end for
9:   end for
10:  if  $L_{ij} > k$  then
11:     $A_{ij} = 0$ 
12:  end if
13: end for
```

2.5 Restricted Additive Schwarz

A preconditioner system is always solved at least once in a loop of an iterative algorithm. Its solution speed has great influence on the entire solution process. A GPU platform provides hundreds of CUDA cores to complete a parallel task. If we can improve the parallel structure of a preconditioner matrix, the solution process can be accelerated. Cai et al. proposed a Restricted Additive Schwarz method to optimize the parallel structure for a preconditioner, as illustrated by Figure 7. The original matrix A is partitioned into some sub matrices first. By the METIS method mentioned above, we got these rectangular matrices whose pivot blocks are dense and other positions are sparse; see Figure 7-(2). Because

the ILU factorization only needs an approximate factorization result from A , we can remove the sparse entries situated outside the pivot blocks. Analyzed from a graph aspect, the entries in the pivot blocks represent vertices and the entries outside them represent edges. If we remove the edges from the graph by RAS process, the communication among GPUs are ruled out. The remained pivot blocks can be solved in parallel. The improvement of parallel performance leads to an accuracy decrease as we discard some entries. So more iteration times are required to reach a convergence. There is an alternative way named overlap to compensate for the loss of accuracy. As shown by Figure 7-(3), the overlap technique requires each pivot block to include its some layers of neighbor entries into the block matrices to be computed. Extra entries improve the calculation accuracy and reduce the iteration times. But extra entries also have a negative influence on the parallel performance. Parallelization and convergence like a cake. We cannot eat it and have it. This characteristic is reflected in the numerical experiment section. As a multiple-GPU platform has two levels of parallelization, the situation becomes complex. One level is composed by the GPUs. The other level is the CUDA cores on each GPU. Both levels need a partition and a overlap.

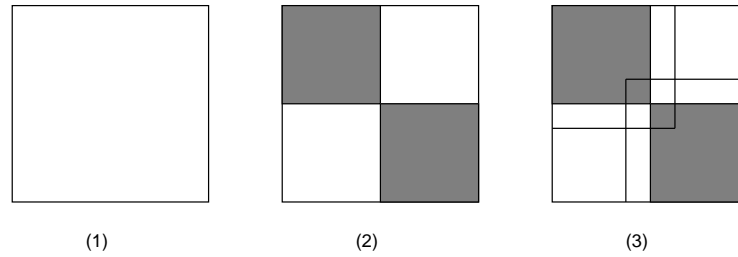


Fig. 7. Restricted Additive Schwarz

2.6 Parallel Triangular Solver

In order to solve L and U on GPUs, we design a parallel triangular solver based on the level schedule method. As an upper triangular system can be easily changed into a lower one, only the lower triangular system is analyzed. The algorithm of a parallel triangular solver is divided into two steps. Each unknown $x(i)$ is assigned a level which is defined by equation (11) in the first step [2]. The second step is the solution process. The triangular problem is solved level by level. All the unknowns in the same level are solved simultaneously. The first level is dependence free. So it is solved at the very first. After the unknowns in the first level is obtained, the second level becomes free and can be solved. This procedure proceeded until all the levels are computed and all the unknowns are solved. A complete algorithm of the level schedule method is given by Algorithm 3.

$$l(i) = 1 + \max_j l(j) \quad \text{for all } j \text{ such that } L_{ij} \neq 0, i = 1, 2, \dots, n, \quad (11)$$

where

- L_{ij} : the (i, j) th entry of L
- $l(i)$: initialized by zeroes
- n : the number of rows

Algorithm 3 Level schedule method for a lower triangular system, $Lx = b$

```

1: Maximal level is n
2: for k = 1 : n do
3:   start = level(k);
4:   end = level(k + 1) - 1;
5:   for i = start: end do
6:     solve the  $i$ th row;
7:   end for
8: end for
```

2.7 Krylov Iterative Algorithms

By now, we have explained the SpMV, vector operations, preconditioner systems and parallel solution process. All these are components of an iterative algorithm. Krylov subspace algorithms contain a series of iterative Algorithms, such as CG (Conjugate Gradient), GMRES (Generalized Minimal Residual), BiCGSTAB (Biconjugate Gradient Stabilized), etc. We have implemented all of them. For instance, an implementation analysis of the BiCGSTAB is shown in the Algorithm 4. All the operations on GPUs are commented. Detailed principle of BiCGSTAB and other Krylov subspace algorithms can be found in [2,3].

2.8 AMG Algorithms

If the coefficient matrix of a system to be solved is symmetric positive definite, an AMG solver should be a better choice. An AMG algorithm has a $L + 1$ levels architecture. The grid of the level is finer with a smaller level number. So the level 0 is the finest level but the level L is the coarsest level. Figure 8 shows the level structure of an AMG solver. An AMG algorithm can be designed as V-cycle, W-cycle or F-cycle. Figure 8 is a V-cycle which has the best acceleration effect on a parallel platform. W-cycle has the worst effect. An AMG process has two phases. The first one is called a setup phase in which the coarser grids, the smoothers, the restriction and prolongation operators are all established. The second one is the solution phase in which the multiple-levels system is solved.

Algorithm 4 BiCGSTAB algorithm

```

1:  $\mathbf{r}_0 = \mathbf{b} - A\mathbf{x}_0$ ;  $\mathbf{x}_0$  is an initial guess vector ▷ SpMV; Vector update
2: for  $k = 1, 2, \dots$  do
3:    $\rho_{k-1} = (\mathbf{r}_0, \mathbf{r})$  ▷ Dot product
4:   if  $\rho_{k-1} = 0$  then
5:     Fails
6:   end if
7:   if  $k = 1$  then
8:      $\mathbf{p} = \mathbf{r}$ 
9:   else
10:     $\beta_{k-1} = (\rho_{k-1}/\rho_{k-2})(\alpha_{k-1}/\omega_{k-1})$ 
11:     $\mathbf{p} = \mathbf{r} + \beta_{k-1}(\mathbf{p} - \omega_{k-1}\mathbf{v})$  ▷ Vector update
12:  end if
13:  Solve  $\mathbf{p}^*$  from  $M\mathbf{p}^* = \mathbf{p}$  ▷ Preconditioner system
14:   $\mathbf{v} = A\mathbf{p}^*$  ▷ SpMV
15:   $\alpha_k = \rho_{k-1}/(\mathbf{r}_0, \mathbf{v})$  ▷ Dot product
16:   $\mathbf{s} = \mathbf{r} - \alpha_k\mathbf{v}$  ▷ Vector update
17:  if  $\|\mathbf{s}\|_2$  is satisfied then ▷ Dot product
18:     $\mathbf{x} = \mathbf{x} + \alpha_k\mathbf{p}^*$  ▷ Vector update
19:    Stop
20:  end if
21:  Solve  $\mathbf{s}^*$  from  $M\mathbf{s}^* = \mathbf{s}$  ▷ Preconditioner system
22:   $\mathbf{t} = A\mathbf{s}^*$  ▷ SpMV
23:   $\omega_k = (\mathbf{t}, \mathbf{s})/\|\mathbf{t}\|^2$  ▷ Dot product
24:   $\mathbf{x} = \mathbf{x} + \alpha_k\mathbf{p}^* + \omega_k\mathbf{s}^*$  ▷ Vector update
25:   $\mathbf{r} = \mathbf{s} - \omega_k\mathbf{t}$  ▷ Vector update
26:  if  $\|\mathbf{r}\|_2$  is satisfied or  $\omega_k = 0$  then ▷ Dot product
27:    Stop
28:  end if
29: end for

```

As a coarser grid has much smaller dimension size compared to its neighbor finer grid, a problem on a coarser grid is easier to be solved. A restriction operation is used for transferring the problem from a finer level to a coarser level. After the problem on the coarser grid is solved, a prolongation operator is used to transfer the solution back to a finer grid. On level l , let A_l be the system matrix, R_l be the restriction operator and P_l be the prolongation operator. S_l is the pre-smoother and T_l is the post-smoother. An example AMG algorithm for V-cycle can be designed as Algorithm 5.

We have developed the AMG solver with a series of smoothers, coarsening operators and prolongation operators. The smoothers include damped Jacobi and weighted Jacobi, etc. The coarsening operator RS and the prolongation operator RSSTD are proposed by Ruge and Stüben [7,8]. The CLJP coarsening operator is proposed by Cleary et al. [12,13].

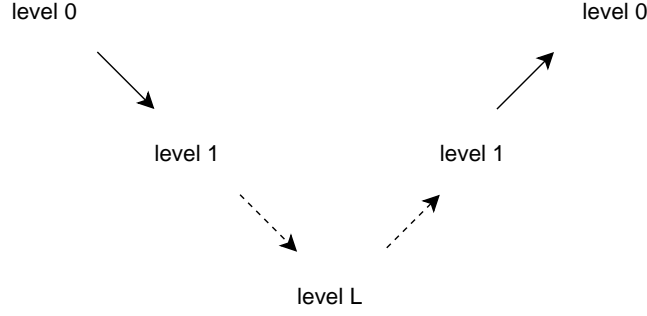


Fig. 8. Structure of AMG solver.

Algorithm 5 AMG V-cycleRequire: $0 \leq l < L$

```

if ( $l < L$ ) then
   $\mathbf{x}_l = S_l(\mathbf{x}_l, A_l, \mathbf{b}_l)$                                 ▷ Pre-smoothing
   $\mathbf{r} = \mathbf{b}_l - A_l \mathbf{x}_l$ 
   $\mathbf{b}_{l+1} = R_l \mathbf{r}$                                        ▷ Restriction
  amg_solve( $l + 1$ )                                         ▷ Recursion
   $\mathbf{x}_l = \mathbf{x}_l + P_l \mathbf{x}_{l+1}$                          ▷ Prolongation
   $\mathbf{x}_l = T_l(\mathbf{x}_l, A_l, \mathbf{b}_l)$                          ▷ Post-smoothing
else
   $\mathbf{x}_l = A_l^{-1} \mathbf{b}_l$ 
end if

```

3 Numerical Experiments

A series of numerical experiments are designed to test our algorithms. We use the speedup to measure the parallel acceleration on GPUs. It is calculated by the ratio of the CPU sequential running time to the GPU parallel running time of the same algorithm. The development environment parameters are listed in Table 1.

Table 1. Experiment environment parameters

Parameter	Value
CPU	Intel Xeon X5570
GPU	NVIDIA Tesla C2050/C2070
Operating System	CentOS X86_64
CUDA Toolkit	5.1
GCC	4.4
CPU codes compilation	-O3 option
float point number precision	double

3.1 SpMV

Table 2 gives the properties of matrices used for SpMV test. *3D_Poisson* is from a three-dimensional Poisson equation. Its dimension is $150 \times 150 \times 150$. The other matrices are all downloaded from a matrix market provided by the University of Florida [24].

Table 2. Matrices for SPMV

Matrix	# of Rows	Nonzeros	NNZ/N	Mb(CSR)
ESOC	327,062	6,019,939	18	70
af_shell8	504,855	9,042,005	18	105
tmt_sym	726,713	2,903,837	4	36
ecology2	999,999	2,997,995	3	38
thermal2	1,228,045	4,904,179	4	61
Hook_1498	1,498,023	30,436,237	20	354
G3_circuit	1,585,478	4,623,152	3	59
kkt_power	2,063,494	7,209,692	3	90
memchip	2,707,524	13,343,948	5	163
3D_Poisson	3,375,000	23,490,000	7	282
Freescall1	3,428,755	17,052,626	5	208
cage15	5,154,859	99,199,551	19	1155

The speedup of SpMV on a single GPU is collected in Table 3. Three matrix formats are tested for each matrix. We can see that most of the speedup for HEC format are over 10 and the highest speedup can reach 18. The algorithm on GPUs has good parallel acceleration performance. Figure 9 makes a comparison of different matrix formats. The number of nonzero entries per row is written in the brackets after each matrix name. We can see that the HEC format represented by the red curve shows better performance than the other two formats. From the Figure, the matrices with relative larger NNZ/N have a lower speedup.

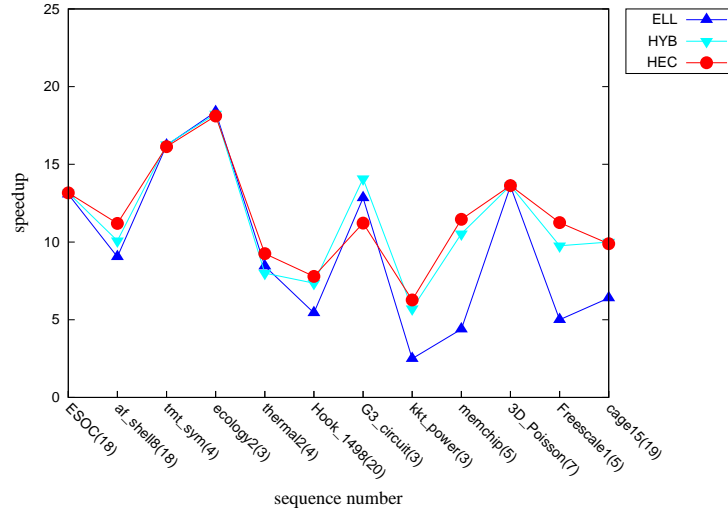
3.2 BiCGSTAB with ILU(K)

In this experiment, we use the BiCGSTAB algorithm with an ILU(k) preconditioner to test our Krylov algorithms. The testing matrix is from a three-dimensional Poisson equation whose dimension is 3,375,000 ($150 \times 150 \times 150$). It has 23,490,000 nonzero entries and about 7 nonzero entries per row. Table 4 collects the running results. There are six parameter combinations which are numbered in the *Seq No.* column. The *Outer RAS* and *Inner RAS* represent the outer layer partition numbers and inner layer partition numbers based on the RAS technique. The number of GPUs employed is equal to the *Outer RAS*. The outer and inner overlap layers are listed in the *Outer overlap* and *Inner overlap* columns, respectively. These parameters form various parameter combinations.

As all the data sections have a similar data tendency, we take the first data section as an sample analysis, where the outer RAS, the inner RAS, the ouer

Table 3. SPMV speedup for different matrix formats

Matrix	ELL	HYB	HEC
ESOC	13.08	13.16	13.16
af_shell8	9.05	10.08	11.20
tmt_sym	16.23	16.27	16.14
ecology2	18.38	18.24	18.11
thermal2	8.45	8.00	9.25
Hook_1498	5.44	7.35	7.79
G3_circuit	12.84	14.08	11.22
kkt_power	2.49	5.71	6.27
memchip	4.39	10.53	11.46
3D_Poisson	13.60	13.63	13.63
Freescall1	5.00	9.76	11.25
cage15	6.40	10.00	9.89

**Fig. 9.** SpMV speedup curves

overlap and the inner overlap are 1, 8, 0 and 0, respectively. The speedup reaches 8.97 when the k level is set to 0. As k goes up from 0 to 3, the speedup goes down from 8.97 to 4.75 in a general data tendency. That is because more fill-in entries are imported by a higher k . These entries contribute to improve the calculation accuracy. So the iteration is saved and goes down from 45 to 33. However, it goes back to 42 when k is 2. That might be caused by the matrix pattern which has also great influence on the performance.

Figure 10 shows a comparison of the combinations. As the outer RAS increases from 1 to 4 and then the inner RAS increases from 8 to 1024, the parallel performance is improved gradually and the curves have a growth tendency.

Table 4. GMRES with ILU(k) for 3D_Poisson (RAS)

Seq No.	Outer RAS	Inner RAS	Outer overlap	Inner overlap	ILU(k) level k	CPU time (second)	GPU time (second)	Speedup	Iteration
1	1	8	0	0	0	16.36	1.82	8.97	45
					1	12.25	1.56	7.86	30
					2	15.75	3.95	3.99	42
					3	16.26	3.42	4.75	33
2	2	8	0	0	0	15.29	1.07	14.30	46
					1	14.64	1.18	12.41	36
					2	18.55	2.66	6.96	43
					3	16.94	2.80	6.05	36
3	3	8	0	0	0	16.57	0.82	20.28	46
					1	14.51	1.07	13.59	39
					2	18.32	2.53	7.25	44
					3	17.85	2.66	6.71	38
4	4	8	0	0	0	17.13	0.62	27.84	44
					1	13.92	0.81	17.14	34
					2	18.15	2.05	8.87	39
					3	17.51	2.47	7.08	38
5	4	128	0	0	0	16.59	0.62	26.96	48
					1	16.91	0.66	25.62	40
					2	20.53	1.50	13.72	51
					3	20.36	1.56	13.02	45
6	4	1024	0	0	0	18.98	0.67	28.33	55
					1	19.37	0.72	27.03	47
					2	21.77	1.39	15.63	58
					3	22.47	1.27	17.74	46

Obviously, a lower k has a better parallel performance. The convergence performance is reflected by the Figure 11. With the sequence number increases, the parallel performance increases but the convergence performance decreases. Thereby more iteration times are needed to reach a convergence. We can see that high iteration times are needed for $k = 0$ because it has high speedup.

As we mentioned, the overlap technique is used for compensating for the loss of calculation accuracy. Higher overlaps are supposed to use smaller iteration. But the speedup is supposed to decrease as more entries are introduced by the overlap. The results of different overlapping configurations are collected in Table 5.

The combination one has the highest speedup 27.82 and iteration 44. Its acceleration performance is the best but convergence performance is the worst. The combination four has an opposite effect with both the outer overlap and inner overlap set to 1. Its speedup is 23.95 and iteration is 38. If only the outer overlap or the inner overlap is set to 1, the results have an intermediate effect.

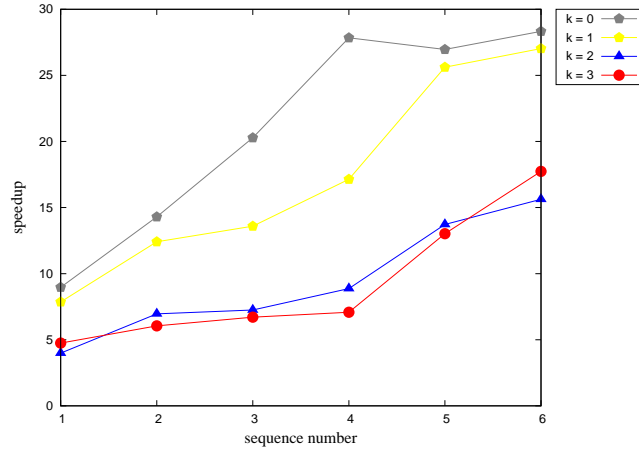


Fig. 10. Speedup for 3D_Poisson

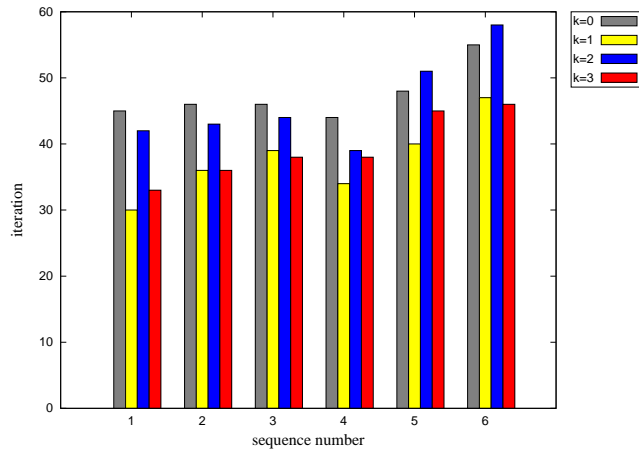


Fig. 11. Iteration for 3D_Poisson

Table 5. GMRES with ILU(k) for 3D_Poisson (overlap)

Seq No.	Outer RAS	Inner RAS	Outer overlap	Inner overlap	ILU(k) level k	CPU time (second)	GPU time (second)	Speedup	Iteration
1	4	8	0	0	0	17.07	0.61	27.82	44
2	4	8	1	0	0	15.91	0.70	22.70	43
3	4	8	0	1	0	15.43	0.60	25.78	41
4	4	8	1	1	0	15.04	0.63	23.95	38

3.3 AMG

Two matrices, *ecology2* and *3D_Poisson*, are employed in the AMG algorithm testing. The *ecology2* is a positive definite matrix derived from a circuit theory applied to animal/gene flow. It has 999,999 rows and 2,997,995 nonzero entries. The NNZ/N is 3. The *3D_Poisson* has an dimension of 125,000 ($50 \times 50 \times 50$) and 860,000 nonzero entries. Its NNZ/N is about 7. We set the maximal level to 8 and the pre-smoothing and post-smoothing both to 3. The V-cycle is employed. Table 6 and 7 collect the running results for *ecology2* and *3D_Poisson*, respectively. Two type of coarsening strategies Ruge- Stüben (RS) and CLJP are used. Two types of interpolations, the standard RS (RSSTD) and direct (RSD), are used. Four types of smoothers are tested. They are the damped Jacobi (dJacobi), weighted Jacobi (wJacobi), Chebyshev polynomial smoothers (Chev) and Gauss-Seidel (GS).

Table 6. AMG for *ecology2*

Seq No.	Coarsening strategy	Interpolation	Smoother	CPU time (second)	GPU time (second)	Speedup	Iteration
1	CLJP	RSD	dJacobi	1.30	0.17	7.57	3
2	CLJP	RSD	Chev	4.92	0.50	9.78	11
3	RS	RSD	dJacobi	0.82	0.11	7.71	3
4	RS	RSSTD	wJacobi	0.86	0.12	7.07	3
5	RS	RSSTD	GS	0.46	0.99	0.46	1

The dJacobi, wJacobi and Chev are all developed based on the SpMV and vector operations. As we have completed the favorable parallel realization of them, these smoothers have good speedup for *ecology2*, which are over 7. When the CLJP and RSD are used, the speedup reaches to the maximal value 9.78. If we select the GS smoother, the speedup is 0.46 which is very low. That means the running time on a GPU is even longer than that on a CPU. The purpose of acceleration on a GPU fails. Although the GS has the worst parallel performance, it has the best convergence performance and only once iteration is needed. So it is better to develop an AMG algorithm with the GS on a CPU. This also shows there is a contradictory effect between acceleration and convergence performance. Our experiment results show that the dJacobi, wJacobi and Chev are suitable for GPU computation while the GS is suitable for CPU.

The *3D_Poisson* has worse acceleration results than the *ecology2* has; shown by Table 7. Different matrices has different nonzero patterns which have great influence on the computing performance. The algorithm on GPU has an acceleration effect for the smoothers of dJacobi, wJacobi and Chev. The combination three with the RS, RSD, dJacobi has the highest speedup 4.71. However, a very poor speedup 0.06 is obtained for the smoother GS. This result is similar to that of the matrix *ecology2*. The GS is not suitable for GPU computing is proved again.

Table 7. AMG for 3D_Poisson

Seq No.	Coarsening strategy	Interpolation	Smoother	CPU time (second)	GPU time (second)	Speedup	Iteration
1	CLJP	RSD	dJacobi	1.64	0.65	2.54	8
2	CLJP	RSD	Chev	1.86	1.13	1.64	8
3	RS	RSD	dJacobi	0.25	0.05	4.71	7
4	RS	RSSTD	wJacobi	0.46	0.13	3.61	9
5	RS	RSSTD	GS	0.28	4.53	0.06	4

4 Conclusion

We have developed the Krylov and AMG linear solvers on GPUs. The SpMV algorithm can be accelerated over 10 times faster on a single GPU against a CPU for most large-scale sparse matrices. Our preconditioned Krylov subspace algorithms have favorable speedups on GPUs. When four GPUs are employed and the inner RAS is set to 1024, the BiCGSTAB with ILU(0) can be sped up to 28 times faster. Our AMG solver shows good parallel performance for dJacobi, wJacobi and Chev smoothers. The numerical experiments verify that a contradictory effect exists between the performance of convergence and acceleration in many cases.

Acknowledgments. The support of Department of Chemical and Petroleum Engineering, University of Calgary and Reservoir Simulation Research Group is gratefully acknowledged. The research is partly supported by NSERC/AIEES/Foundation CMG, AITF iCore, IBM Thomas J. Watson Research Center, and the Frank and Sarah Meyer FCMG Collaboration Centre for Visualization and Simulation. The research is also enabled in part by support provided by WestGrid (www.westgrid.ca) and Compute Canada Calcul Canada (www.computecanada.ca).

References

1. Z. Chen, *Reservoir Simulation: Mathematical Techniques in Oil Recovery*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 77, SIAM, Philadelphia, 2007.
2. Y. Saad, *Iterative methods for sparse linear systems (2nd edition)*, SIAM, 2003.
3. R. Barrett, M. Berry, T. F. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine and H. Vander Vorst, *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods, 2nd Edition*, SIAM, 1994.
4. X. Hu, W. Liu, G. Qin, J. Xu, Y. Yan, C. Zhang, *Development of A Fast Auxiliary Subspace Pre-conditioner for Numerical Reservoir Simulators*, SPE Reservoir Characterisation and Simulation Conference and Exhibition, Abu Dhabi, UAE, SPE-148388-MS, 9-11 October 2011.
5. Z. Chen, G. Huan, and Y. Ma, *Computational Methods for Multiphase Flows in Porous Media*, in the Computational Science and Engineering Series, Vol. 2, SIAM, Philadelphia, 2006.

6. H. Liu, K. Wang, Z. Chen, and K. E. Jordan, *Efficient Multi-stage Preconditioners for Highly Heterogeneous Reservoir Simulations on Parallel Distributed Systems*, SPE-173208-MS, SPE Reservoir Simulation Symposium Held in Houston, Texas, USA, 23-25 February 2015.
7. K. Stüben, *A review of algebraic multigrid*, Journal of Computational and Applied Mathematics Volume 128, Issues 1-2, 2001, 281–309.
8. J.W. Ruge and K. Stüben, *Algebraic multigrid (AMG)*, in: S.F. McCormick (Ed.), Multigrid Methods, Frontiers in Applied Mathematics, Vol. 5, SIAM, Philadelphia, 1986.
9. A. Brandt, S.F. McCormick and J. Ruge, *Algebraic multigrid (AMG) for sparse matrix equations* D.J. Evans (Ed.), Sparsity and its Applications, Cambridge University Press, Cambridge, 1984, 257–284.
10. C. Wagner, *Introduction to Algebraic Multigrid*, Course notes of an algebraic multigrid course at the University of Heidelberg in the Wintersemester, 1999.
11. P. S. Vassilevski, *Lecture Notes on Multigrid Methods*, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2010.
12. R. Falgout, A. Cleary, J. Jones, E. Chow, V. Henson, C. Baldwin, P. Brown, P. Vassilevski, and U. M. Yang, *Hypre home page*, 2011. <http://acts.nersc.gov/hypre>
13. A. J. Cleary, R. D. Falgout, V. E. Henson, J. E. Jones, T. A. Manteuffel, S. F. McCormick, G. N. Miranda, and J.W. Ruge, *Robustness and Scalability of Algebraic Multigrid*, SIAM J. Sci. Comput., 21, 2000, 1886–1908.
14. Y. Gao, S. Iqbal, P. Zhang, M. Qiu, *Performance and Power Analysis of High-Density Multi-GPGPU Architectures: A Preliminary Case Study*, 2015 IEEE 17th International Conference on High Performance Computing and Communications (HPCC-ICESS-CSS 2015), New York, USA, August 24-26, 2015.
15. NVIDIA Official Website, <http://www.nvidia.com/object/tesla-servers.html>.
16. Fujitsu Official Website, <http://techcommunity.ts.fujitsu.com/en/client-computing-devices-2/d/uid-5911b36b-324b-fc23-45fa-2438e4c546f3.html>.
17. G. Karypis and V. Kumar, *A Fast and Highly Quality Multilevel Scheme for Partitioning Irregular Graphs*, SIAM Journal on Scientific Computing, 20(1), 1999, pp. 359-392.
18. X.-C. Cai and M. Sarkis, *A Restricted Additive Schwarz Preconditioner for General Sparse Linear Systems*, SIAM J. Sci. Comput., 21, 1999, pp. 792-797.
19. H. Liu, S. Yu, Z. Chen, B. Hsieh and L. Shao, *Sparse Matrix-vector Multiplication on NVIDIA GPU*, International Journal of Numerical Analysis & Modeling, Series B, Volume 3, 2012, No. 2, pp. 185-191.
20. H. Liu, S. Yu, Z. Chen, B. Hsieh and L. Shao, *Parallel Preconditioners for Reservoir Simulation on GPU*, SPE 152811-PP, SPE Latin American and Caribbean Petroleum Engineering Conference held in Mexico City, Mexico, 16-18 April 2012.
21. R. Li and Y. Saad, *GPU-accelerated Preconditioned Iterative Linear Solvers*, Technical Report Umsi-2010-112, Minnesota Supercomputer Institute, University of Minnesota, Minneapolis, MN, 2010.
22. R. Grimes, D. Kincaid, and D. Young, *ITPACK 2.0 User's Guide*, Technical Report CNA-150, Center for Numerical Analysis, University of Texas, August 1979.
23. N. Bell and M. Garland, *Implementing Sparse Matrix-vector Multiplication on Throughput-oriented Processors*, Proc. Supercomputing, November 2009, pp. 1-11.
24. T. A. Davis, *University of Florida Sparse Matrix Collection*, NA digest, 1994, <https://www.cise.ufl.edu/research/sparse/matrices/>.

